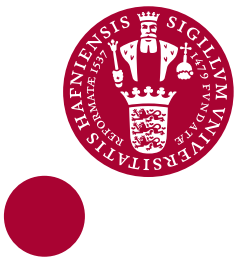

Darkness in the Human Gene and Protein Function Space

Widely Modest or Absent Illumination by the Life Science Literature and the Trend for Fewer Protein Function Discoveries Since 2000

Sinha, Swati; Eisenhaber, Birgit; Jensen, Lars Juhl; Kalbuaji, Bharata; Eisenhaber, Frank

Published in:
Proteomics

DOI:
10.1002/pmic.201800093

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

RESEARCH ARTICLE

Gene Function Discovery

Proteomics

www.proteomics-journal.com

Darkness in the Human Gene and Protein Function Space: Widely Modest or Absent Illumination by the Life Science Literature and the Trend for Fewer Protein Function Discoveries Since 2000

*Swati Sinha, Birgit Eisenhaber, Lars Juhl Jensen, Bharata Kalbuaji, and Frank Eisenhaber**

The mentioning of gene names in the body of the scientific literature 1901–2017 and their fractional counting is used as a proxy to assess the level of biological function discovery. A literature score of one has been defined as full publication equivalent (FPE), the amount of literature necessary to achieve one publication solely dedicated to a gene. It has been found that less than 5000 human genes have each at least 100 FPEs in the available literature corpus. This group of elite genes (4817 protein-coding genes, 119 non-coding RNAs) attracts the overwhelming majority of the scientific literature about genes. Yet, thousands of proteins have never been mentioned at all, ≈2000 further proteins have not even one FPE of literature and, for ≈4600 additional proteins, the FPE count is below 10. The protein function discovery rate measured as numbers of proteins first mentioned or crossing a threshold of accumulated FPEs in a given year has grown until 2000 but is in decline thereafter. This drop is partially offset by function discoveries for non-coding RNAs. The full human genome sequencing does not boost the function discovery rate. Since 2000, the fastest growing group in the literature is that with at least 500 FPEs per gene.

1. Introduction

The biomolecular mechanisms are the holy grail of modern biology and medicine. It is the mechanistic understanding that opens a rational way to influence processes in living systems with predictable outcomes.[1] A complete catalogue of functions associated with genes and other genomic regions is a necessary first step toward mechanistic insight into biological systems. As researchers involved in the interpretation of omics data from various screens in terms of biological and medical implications on a daily basis and over many years, we know that, regularly, functionally uncharacterized genes pop up in those studies and the status of many of those genes has remained unchanged ever since.[2,3]

Notably, the biological function of a gene, protein or non-coding RNA is a hierarchical concept including aspects of molecular, cellular, and phenotypic function.[4] Therefore, many scientific efforts (and subsequent scientific publications) are necessary to explore and to report all those features. An incomplete list involves genetic screens, analysis of mutations in suitable model organisms, omics studies, cell biology work, structural biology research, sequence-analytic comparisons (for example, as in ref. 5), clinical applications, etc. The critical point is usually the discovery of biomolecular mechanisms, of enzymatic and binding activities, of critical conformational changes as well as of interactions with other biomacromolecules and small compounds. Beyond qualitatively establishing relationships, the quantification of turnover rates, binding affinities, fluxes, etc. represents another formidable challenge.

With this publication, we wish to provide quantitative assessments of discovery rates of genomically encoded biological functions in various historical periods of time. As lots of the scientific literature—abstracts and full-text versions of academic papers, patents, and other related documents—are publicly available, it is possible to assess first and repeated occurrences of gene, protein, and non-coding RNA names in the literature corpus as a function of publication date and, thus, indirectly evaluate the level of

Dr. S. Sinha, Dr. B. Eisenhaber, B. Kalbuaji[+], Dr. F. Eisenhaber
Bioinformatics Institute (BII)
Agency for Science and Technology (A*STAR)
Matrix, 138671, Singapore
E-mail: franke@bii.a-star.edu.sg

Dr. L. J. Jensen
Novo Nordisk Foundation Center for Protein Research
Faculty of Health and Medical Sciences
University of Copenhagen
DK-2200 Copenhagen, Denmark

Dr. F. Eisenhaber
School of Computer Science and Engineering (SCSE)
Nanyang Technological University (NTU)
637553, Singapore

[+]Present address: Department of Computer Science, School of Computing, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152–8552, Japan

DOI: 10.1002/pmic.201800093

functional insight for those genomically defined entities and changing trends of annual function discovery rates. About 17 years have gone since the presentation of the draft of the human genome[6,7]; so, it would also be interesting to see whether the availability of the full human genome sequence could be correlated with some changes (expectedly, an increase) in the annual rate of discovery of functions associated with certain genomic regions.

Significance Statement

It is generally believed that full human genome sequencing was a watershed event in human history that boosted biomedical research, biomolecular mechanism discovery, and life science applications. At the same time, researchers in the field of genome annotation see that there is a persisting, substantial body of functionally insufficiently or completely not characterized genes (for example, ≈10 000 protein-coding genes in the human genome) despite the availability of full genome sequences. A survey of the biomedical literature shows that the number of reported new protein functions had been steadily growing until 2000 but the trend reversed to a dramatic decline thereafter. The fastest growing set of genes in the last decade is that with 500 or more full publication equivalents, i.e., the genes that are well characterized anyhow. At the same time, the annual amount of life science publications doubled between 2000 and 2017. There are no apparent scientific or financial reasons for the decline in biomolecular mechanism discovery; probably, current instruments of science funding do not direct or even discourage researchers to go after the difficult problems of gene function discovery.

2. Experimental Section

For the goal of this work, it was necessary to associate names of human genes, proteins, and non-coding RNAs ("genomically determined named entities"), with identifiers ("publication IDs") of papers among the published scientific literature, patents, etc. For this purpose, previously applied approaches were reused.[8,9] The text mining technology aimed at fractional counting of genomic entities' mentions required (i) a list of entity names and their synonyms as well as (ii) the body of the available, published life science literature. In both cases, the data was desirably as complete as possible.

2.1. Named Entity Recognition and Text Corpus Construction

The recognition of protein and gene names (abbreviated as NER, named entity recognition below) in scientific texts built upon the text-mining pipeline applied for generating the STRING database.[9] This highly efficient and flexible NER engine was implemented in C++ and had been described in full detail elsewhere.[10] Briefly, the updated entity name dictionaries and the underlying human genome annotation from STRING version v10.5[11,12] and, additionally for non-coding RNAs, from RAIN were used.[13] The dictionary merges synonym information from multiple sources, including the Ensembl[14] and UniProt[15] databases. An explicit rule system described in ref. 16, which combines sets of regular expressions and a list of blocked names, was applied to suppress the recognition of entity names in target texts when the respective words were frequently used to mean something else, for example, in the case of certain acronyms and common English words.

To construct a literature corpus, all articles were first downloaded from the PubMed Central (PMC) Open Access Subset (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). Next, PubMed was queried for additional articles for which full-text was freely available online and automatically downloaded these in PDF format from the publisher's website when possible. AbiWord (http://www.abiword.org/) was used to convert them to HTML format. Because the text was usually extracted from PDF files, the text might be subjected to formatting artifacts. The resulting combined corpus consisted of >2 million full-text articles. All text sources were subsequently converted to the format required by the NER software. Despite this effort, presently there was no access so that the full-text articles could be mined for the majority of Medline entries.

Thus, the literature corpus was extended with these abstracts, which were extracted from a local copy of Medline (http://www.nlm.nih.gov/databases/journal.html) and converted to the same format as the full-text articles. It needs to be emphasized that named entity recognition delivered richer results from full-text sources than from abstracts only and this has implications for finding literature evidence about genes, for example, for protein–protein and protein–cellular compartment associations but especially for links between diseases and genes.[17]

The literature corpus was complemented with a second corpus of USPTO patent texts, which were downloaded from Google Books (http://www.google.com/googlebooks/uspto-patents-grants-text.html). As the file format of these had changed over the years, several parsers were developed to convert them all into the same unified format as the literature corpus. Most notably, all patents prior to 2002 were scanned and converted to plain text through optical character recognition (OCR); these were, thus, also could be subject to OCR errors.

Thus, the whole procedure compounds PDF and OCR transformation errors with those of natural language processing for named entity recognition (see ref. 17 for extensive discussion). For example, entity names absent in the list would not be recognized in the search. Literature documents absent in the study of the literature corpus would reduce the possible counts for some named genomic entities. Most likely, not all genomic entity names would be found (though the number of false negatives was difficult to estimate). Also, certain false positive assignments would be made. Thus, the data was not useful to make judgments for a specific gene (such as date of first publication or the exact total number of publications about it). Yet, the data should be good enough to assess statistical trends with regard to publication rates about gene functions as the error rate was to be expected roughly the same for every year. Of course, the equal probability of false-positive/false-negative errors over the years was just an assumption that was reasonable although axiomatic.

The named entity recognition within the natural language processing approach, as applied in this work, would recognize the

mentioning of the human gene (or protein, ncRNA, etc.) but it would also hit cases when the text was about a gene with the same name in mouse, rat, etc. It was assumed in this work that the article reported about the function of an orthologous (or closely homologous) gene in another model organisms, it was relevant for the function of the human gene and this publication should be counted for this purpose even if the human ortholog was not mentioned at all.

At the same time, the coarseness of the synonym lists that were at the disposal did not allow to systematically track individual splicing variants or protein isoforms. So, they were all lumped together under a single gene name.

There were also trends of name agglomeration that mentioning one gene was typically followed by naming other interacting genes in the same text such as co-expressed transcripts or subunits in protein complexes. In the concept of fractional counting, these groups of genes were thus assigned coherently growing score with each additional publication of this kind.

2.2. Fractional Counting of Entity Names and Determination of Full Publication Equivalents

A document could mention multiple proteins without pertaining equally much to all of them. To address this, a fractional counting scheme[8] was used in which each paper that mentioned at least one protein contributed a total count of 1, which was distributed across the mentioned proteins relative to how many times each of them was mentioned. The total fractional count f_i for protein or gene i was thus:

$$f_{\mathrm{i}} = \sum_{\mathrm{j}\in \mathrm{D}} \frac{n_{\mathrm{ij}}}{n_{\cdot \mathrm{j}}}$$

where D is the document set, n_{ij} is the number of times protein or gene i as mentioned in document j, and n_{ij} is total number of mentions of any protein or gene in document j.

A master file was generated where each line contained a genomic entity name, a publication identifier, the publication date, and the fractional count associated with that genomic entity name. From this source, it was possible to assess the amount of literature published about a given genomic entity (the literature score) in periods of time by summing up the respective fractional counts for publications in the years considered. A literature score of one was defined as full publication equivalent (FPE), the amount of literature necessary to achieve one publication solely dedicated to a single genomic entity (gene, protein, or non-coding RNA). As shown in ref. 8, more publications per named genomic entity strongly correlated with more complete insight into its functional aspects. Thus, further in the text, the number of FPEs per named genomic entity was used as proxy for the level of knowledge about its biological function.

For standard statistical tests, the software "R" and Microsoft Excel were used.

3. Results

3.1. Mapping of Life Science Literature onto Human Genomic Regions: Status End of 2017

Here, we report our results of mapping the life science literature from its early beginning until the end of 2017 (the last completed year as of the day of writing this article) onto the human genome by computing the number of historically accumulated FPEs per named entity. This literature score (the number of FPEs) is then used to deduce insights about the level of functional description and involvement in biomolecular mechanisms for those genes, proteins, and non-coding RNAs.

It is difficult to assess how many FPEs are necessary until all critical aspects of biological function of a genomically determined named entity are fully described. For example in the case of PIG-K (gpi8), it took less than 3 years and only five papers for the discovery of the protein, its location in the endoplasmic reticulum and its role as protease subunit in the transamidase complex responsible for attaching GPI lipid anchors to substrate proteins.[4,18–21] One of its most tightly bound protein partners, the subunit GAA1/GpAA1 was discovered also in 1995.[22] But its function as synthetase of the peptide bond linking the GPI lipid anchor and the C-terminus of the substrate protein became clear only ≈20 years later (in 2014[23]) and, in between, there were about another 15 papers dealing with various aspects of GAA1/GpAA1's functional significance.

Therefore, we explore different FPE ranges for the literature score S of a named genomic entity (the ranges are $0 < S < 1$, $1 \leq S < 10$, $10 \leq S < 20$, …, $100 \leq S < 500$, $500 \leq S$; see **Table 1** for protein-coding genes and **Table 2** for other genes and non-coding RNAs). The trends among them allow us to better understand how the body of literature covers all aspects of function including the biomolecular mechanisms involving the genes and proteins.

The literature body accumulated from 1901 until 2017 studied in this work contains references to 17 824 proteins and 2641 non-coding RNAs (with ≈4.6 million FPEs in total). In pie charts of **Figure 1**, we show how the numbers of named entities and their associated literature scores are spread among FPE ranges. We find that the scientific literature is distributed extremely unevenly for protein-coding and even more so for non-coding genes.

The most mentioned 9% of all proteins (1610 entities, Figure 1A and Table 1) each attracted >500 FPEs and, together, this forms 78% of the total body of the literature (Figure 1C). Among the most studied proteins each with 32 000–145 000 FPEs, we find insulin, serum albumin, p53, tumor necrosis factor (TNF), CD40, pro-opiomelanocortin (the precursor of several peptide hormones such as melanin, endorphin, encephalin, ACTH, etc.), C-reactive protein, renin, and maltase-glucoamylase. These ten proteins alone have ≈450 000 FPEs taken together.

Notably, some items (p53 and TNF) overlap with the list of ten most heavily researched proteins as reported by Dolgin.[24] Other proteins mentioned in ref. 24 also occupy high ranks in our list: IL6 (rank 11), AKT1 (21), VEGFA (23), APOE (28), EGFR (51),

Table 1. Status of mapping the life science literature between 1900 and 2017 onto protein-coding genes. The total number of protein targets, their percentage among the total 17824 mentioned in the literature, their accumulated literature score S measured in FPEs, and their share among the total amount of FPEs for all protein-coding genes are listed for different ranges of FPEs per target. The ranges for the literature score are $0 < S < 1$, $1 \leq S < 10$, $10 \leq S < 20$, ..., $100 \leq S < 500$, and $500 \leq S$.

Range for S	Number of proteins	Percentage of 17 824 proteins	Total literature score for all targets	Percentage of total score
0–1	1997	11.2	836.2	0.02
1–10	4571	25.6	20 379.0	0.44
10–20	1935	10.8	27 957.7	0.60
20–30	1175	6.6	28 863.9	0.62
30–40	857	4.8	29 730.9	0.64
40–50	638	3.6	28 597.8	0.61
50–60	531	3.0	29 090.5	0.62
60–70	352	2.0	22 808.3	0.49
70–80	357	2.0	26 759.0	0.57
80–90	312	1.8	26 456.9	0.57
90–100	282	1.6	26 779.5	0.57
100–500	3207	18.0	73 6109.8	15.75
>500	1610	9.0	3 666 853.2	78.50

Table 2. Status of mapping the life science literature between 1900 and 2017 onto non-protein-coding genes. The total number of protein targets, their percentage among the total 2641 mentioned in the literature, their accumulated literature score S measured in FPEs, and their share among the total amount of FPEs for all protein-coding genes are listed for different ranges of FPEs per target. The ranges for the literature score are $0 < S < 1$, $1 \leq S < 10$, $10 \leq S < 20$, ..., $100 \leq S < 500$, and $500 \leq S$.

Range for S	Number of genes	Percentage of genes	Total literature score for all targets	Percentage of total score
0–1	1357	51.38	520.0	0.81
1–10	816	30.90	2818.3	4.37
10–20	139	5.26	2035.2	3.16
20–30	60	2.27	1511.7	2.34
30–40	42	1.59	1488.9	2.31
40–50	24	0.91	1070.2	1.66
50–60	25	0.95	1380.9	2.14
60–70	16	0.61	1024.8	1.59
70–80	15	0.57	1125.0	1.74
80–90	16	0.61	1377.9	2.13
90–100	12	0.45	1147.4	1.78
100–500	102	3.86	19 459.7	30.17
>500	17	0.64	29 533.2	45.80

ESR1 (72), TGFB1 (119), and MTHFR (129). Each of these targets is described with 5000–31 000 FPEs. The differences in the rankings can be associated with specifics of the selection methodology. In that work, a much smaller body of literature was studied and the association of publications with genes was based on counting at least a single mentioning of their names in the text as full FPE.

Further 18% of all proteins listed (3207) have a literature score S between 100 and 500 FPEs that, taken together, represent another 16% of the literature. Thus, an elite group of slightly less than 5000 especially well-studied proteins (4817 or 27% of all proteins studied during all history of science) attracted 94.2% of all life science publications about proteins and genes encoding them.

At the same time, only 6% of the FPEs cover the known aspects of function for further ≈13 000 entities. For 6439 protein-coding genes and derived proteins (36% of all proteins ever mentioned), there is between 10 and 100 FPEs in the literature. Taken together, this comprises 4.5% of the total publication corpus. Presumably, 10–100 FPEs correspond to some basic level of functional insight reported. For another subset of 1997 proteins (11%), there is not even one full FPE in the literature. For additional 26% of protein entries (4571), the number of FPEs counted is below ten (in average, around five). Thus, the functions of these latter two groups, 37% of all proteins ever mentioned in the literature (in just 0.5% of the literature corpus), are barely known. This is on top of thousands of additional proteins, the sequences of which are known as a result of the human genome

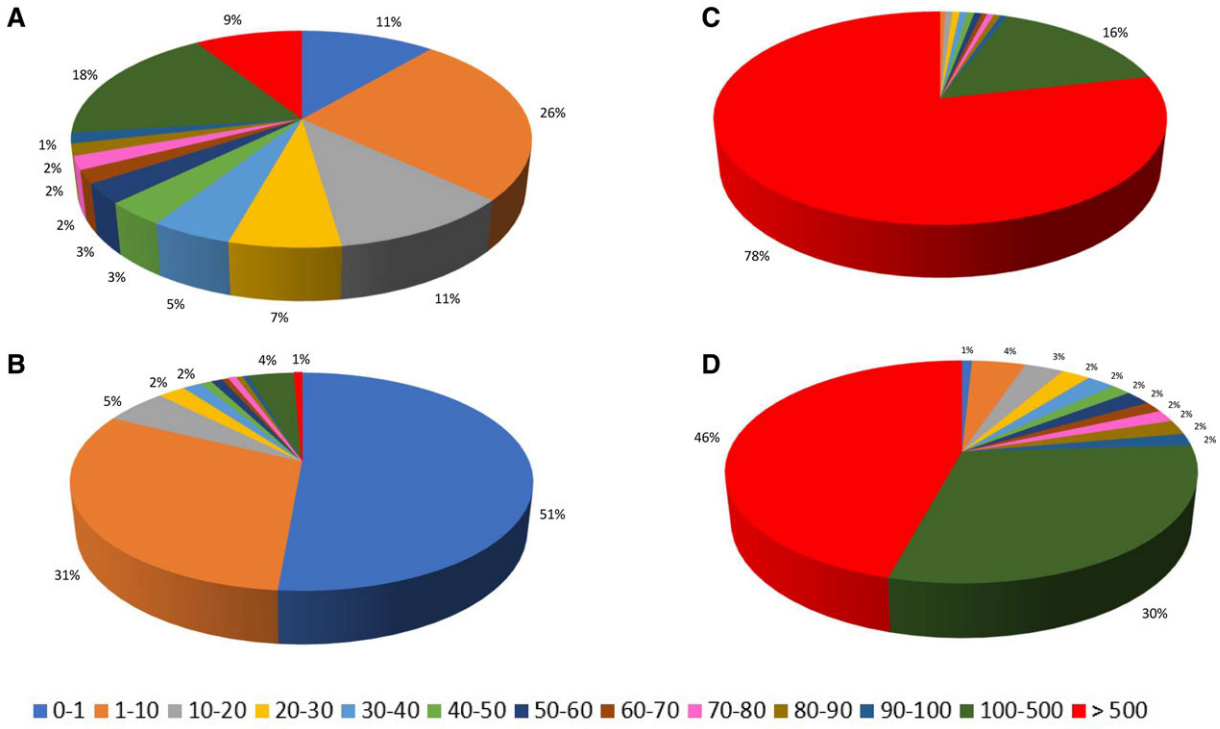

Figure 1. Status of the mapping of life science literature accumulated until 2017 onto the human genome. For various FPE ranges ($0 < FPE < 1$, $1 \leq FPE < 10$, $10 \leq FPE < 20$, $20 \leq FPE < 30$, $30 \leq FPE < 40$, $40 \leq FPE < 50$, $50 \leq FPE < 60$, $60 \leq FPE < 70$, $70 \leq FPE < 80$, $80 \leq FPE < 90$, $90 \leq FPE < 100$, $100 \leq FPE < 500$, $500 \leq FPE$), the distribution of (A) the number of proteins and (B) the number of non-coding RNAs is shown as pie chart. The accumulated FPEs (the total literature score) of the named entities within those FPE brackets is presented in (C) for protein-coding genes and (D) for non-protein-coding genes.

project,[6,7,25] that have never received any attention in a functional study.

Among non-coding genes and RNAs, the situation is even more exaggerated (Figure 1B and D, Table 2). Only 1% of the 2641 non-coding RNAs (17 entities) have more than 500 FPEs, 4% (additional 102) score between 100 and 500 FPEs. Together, these 5% non-coding RNAs (119 entities, usually miRNAs implicated in cancer) harbor 76% of all literature on the topic of non-coding RNA structure and function. At the same time, 51% (1357) of all non-coding RNAs mentioned in the literature have not even one full FPE dedicated to their functional description. The total number of non-coding RNAs with physiological relevance has not even been estimated but, most likely, will exceed the number of protein-coding genes. Thus, the knowledge gap is much larger here.

3.2. Trends in the Discovery of New Protein-Coding Gene and Protein Functions

Obviously, the current level of functional insight about many named genomic entities is not satisfactory, especially since most of the efforts of the scientific community have been directed on studying targets that have received disproportional attention anyhow. Although the actual situation is not desirable, maybe, the trends in function discovery are encouraging with lots of recent progress?

We studied the rate that new genomic entities appeared in the literature at the first time and in which year they crossed certain thresholds of their accumulated FPEs. In the following text, the notion "Tx" ("**T**hreshold **x**", where "x" is a natural number) is used to describe thresholds for a given named entity to cross the literature score threshold with "x" FPEs. So, "T0" denotes at least a single occurrence (score>0) in the scientific literature. "T1" requires at least one full FPE to be accumulated. Accordingly, other thresholds such as T5, T10, T15, T20, T25, T30, T35, T40, T45, T50, T75, T100 and T500 are to be similarly understood.

In **Figure 2**, we illustrate the data about how many protein-coding genes reach a certain level of literature representation in a given year for the first time. Without any rigorous statistical methodology, certain periods are clearly delineated. Between 1900 and 1945, there was almost no function discovery effort reflected in the literature. Although the concept of gene was just in the process of emerging, occasional but important protein function and pathway discoveries such as glycolysis,[26] the tricarboxylic acid cycle[27] or the ATPase activity of myosin[28] have been made.

But thereafter until ≈1975, insight into biomolecular mechanisms improved dramatically with increasing increments of new knowledge essentially every year. This process culminated into a spike of gene function discovery around 1975 with 667 genes mentioned in scientific articles for the first time and 134 genes having accumulated ≥10 FPEs in that year alone.

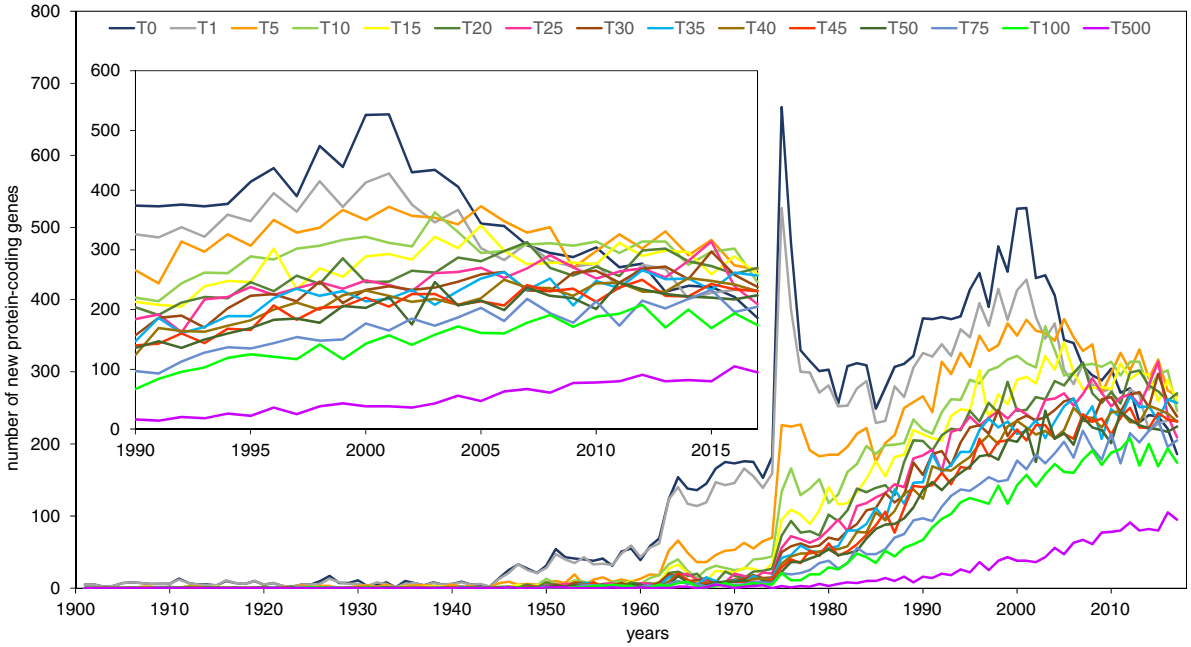

Figure 2. The number of new protein-coding genes in a given year with accumulated FPE score crossing different thresholds the first time. The notion "Tx" (where "x" is a natural number) is used to describe thresholds for a given named entity to cross the literature score threshold with at least "x" FPEs. So, "T0" denotes at least a single occurrence of the gene name (score > 0) in the scientific literature. "T1" requires at least one full FPE to be accumulated. Accordingly, other thresholds such as T5, T10, T15, T20, T25, T30, T35, T40, T45, T50, T75, T100, and T500 are defined respectively. See Supporting Information file 1 for the exact protein-coding gene numbers. The graph from 1990–2017 is zoomed on the left and shown in box.

Despite a ≈50% drop in the annual function discovery rate immediately after the 1975 pulse, the period 1980–2000 saw a steady increase of the protein function discovery rate as evidenced by ever larger numbers of genes crossing all studied literature score thresholds, especially visible from curves showing T0, T1 and T5. At around 2000, about 500 protein names per year appeared in the scientific literature for the first time. And about 300 new proteins crossed the threshold T10 every year.

The outcome for the years after 2000 is nothing but surprising. As a draft of the full human genome sequence was presented to the community in 2001[6,7] and a completed version became available in 2004,[25] the full sequences of essentially all genes became publicly available and a major hurdle towards studying new genomic entities was thought to be removed. Nevertheless, we observe a sharp drop and continuous decrease of the rate that new proteins appear in the literature (curve for T0, 186 new entries in 2017 compared with 526 in 2000) or cross certain FPE thresholds (curves for T1, T5, T10, T15). The rates for medium thresholds (T20-T100) tend to stabilize at a constant (between 200 and 250 new entries). At the same time, the rate for new proteins to cross T500 increases drastically.

This visual impression is confirmed by a rigorous statistical analysis of growth per year in various eras (1945-1974, 1980–1999, and 2000–2017; see **Table 3**). We studied the regression lines approximating the curves in Figure 2. The slope of the regression corresponds to the additional number of proteins that crossed a FPE threshold compared with the previous year; i.e., the slope approximates the average annual change in the discovery rate in the respective era.

During the first time period 1945–1974, there is strong growth for new proteins appearing in the literature for the first time (6.4 additional new items per year compared with the rate in the previous year) and for crossing thresholds T1, T5, T10 and T15. The trend for robust growth is further enhanced during 1980–1999 when there are eight to ten additional new proteins in addition to the rate of the previous year that crosses any of the thresholds T0, T1, T5, …, T50, and T75. The growth rate is smaller for T100 (≈6 additional new proteins per year) and very low for T500 (1-2 additional new proteins per year). In all cases, the P-value of the F-test is highly significant and supporting the growth trend (P-value < 1.e-7 for all entries 1980–1999 and <1.e-4 for all entries 1945–1974 except for T500, see Table 3).

The regression data emphasizes the qualitative change in the time period after 2000. Statistically significant growth rates become largely negative for T0 and T1 but they are negative for T5, T10, and T15, too. P-values for the F-test indicate that, for T15, T20, …, T45, and T50, the discovery rate as measured by the number of genes crossing FPE thresholds in the given year is essentially constant. Some growth is observed for T75 and T100 (2.3–2.5 new proteins per year) and strong, statistically significant growth is only seen for new proteins crossing the T500 threshold (3.7 new proteins per year). By the way, this is the only slope that has increased (from 1.7) compared with the previous era 1980–2000.

It should be noted that time intervals selected (1945–1974, 1980–1999, and 2000–2017) correspond to what we visually perceived as periods with differing dynamics of function discovery. Some variation of boundary years as suggested by a reviewer will

Table 3. Changes in numbers of newly functionally characterized protein-coding genes in three different time periods. The protein-coding gene is considered functionally characterized if its literature score has crossed a minimal "literature score threshold" Tx in a given year the first time (see main text for Tx definitions and also Figure 2 for illustration). For three time periods (1945–1974, 1980–1999, and 2000–2017), we approximate the curves in Figure 3 with regression lines to estimate trends for function discovery. The regressions are characterized by "slope" (change of number of new protein-coding genes per year), R^2 (residual), ρ (correlation coefficient), and its p-value by an F-test.

	1945–1974				1980–1999				2000-2017			
Tx	Slope	R^2	ρ	p-Value	Slope	R^2	ρ	p-Value	Slope	R^2	ρ	P-value
0	6.4	0.8495	0.922	4.91e-13	9.5	0.8078	0.899	7.28e-08	−17.97	0.9026	−0.950	1.67e-09
1	5.6	0.8602	0.927	1.74e-13	8.0	0.7779	0.882	2.73e-07	−11.1	0.8515	−0.923	4.98e-08
5	2.3	0.7811	0.8838	9.72e-11	9.97	0.9028	0.950	1.51e-10	−5.1	0.6523	−0.807	5.06e-05
10	1.3	0.7499	0.866	6.41e-10	9.6	0.9601	0.980	4.8e-14	−2.6	0.3582	−0.598	0.008687
15	0.98	0.6756	0.822	2.56e-08	8.4	0.8814	0.939	9.12e-10	−1.5	0.1637	−0.404	0.0958
20	0.8	0.6705	0.819	3.19e-08	9.4	0.9439	0.972	1.05e-12	0.9	0.0561	0.236	0.3441
25	0.6	0.6776	0.823	2.34e-08	9.3	0.9461	0.973	7.26e-13	0.5	0.0137	0.116	0.644
30	0.5	0.6230	0.789	2.17e-07	9.3	0.9409	0.970	1.67e-12	1.7	0.2719	0.521	0.02647
35	0.5	0.5395	0.735	3.82e-06	9.7	0.9644	0.982	1.7e-14	2.1	0.3438	0.586	0.01054
40	0.4	0.5399	0.735	3.77e-06	9.2	0.9733	0.987	1.31e-15	1.4	0.3268	0.571	0.01319
45	0.4	0.4743	0.689	2.58e-05	8.9	0.9459	0.973	7.55e-13	1.3	0.2801	0.529	0.0239
50	0.3	0.5031	0.709	1.14e-05	8.7	0.9741	0.987	9.8e-16	1.1	0.1087	0.329	0.1815
75	0.3	0.6105	0.781	3.46e-07	7.3	0.9618	0.981	3.29e-14	2.3	0.3994	0.631	0.0049
100	0.2	0.4628	0.680	3.53e-05	6.2	0.9187	0.958	3.0e-11	2.5	0.4842	0.695	0.00134
500	0.1	0.2538	0.504	0.004537	1.7	0.8232	0.9073	3.41e-08	3.7	0.9029	0.950	1.61e-09

deliver somewhat changed slopes compared to those in Table 3 but the message will remain the same: there was a substantial, positive dynamics of the discovery rate before 2000 that disappeared after the beginning of the new millennium.

It was already known[17] that, in average, the number of mentioning of gene names (not just new ones) in the literature had a trend towards moderate growth from the early beginnings until ≈1975 (from a low number 1–3 to about ≈9 per year and per gene name). This number remained about constant until ≈2000 and, then, it increased drastically towards about 22 in 2017 (see Figure S4, Supporting Information in,[17] graph at the bottom). It is notable that the time sections clearly distinguishable in Figure S4 coincide with the periods determined in this work.

Two factors influence the average number of gene mentioning: (i) the total number of publications with gene names (that is generally increasing and driving up the average) and (ii) the rate of introducing new gene names in the literature (that, if increasing, will tend to reduce the average). The slow or zero growth of the average gene mentioning until 2000 indicates that much of the expansion of the scientific literature was about describing new biomolecular entities and their functions and not about reporting news about previously mentioned genes. Yet, it appears that, after 2000, the average mentioning of gene names amplified drastically because of the combination of exponential growth of the scientific article number (see Figure S2, Supporting Information in[17]) and of zero or negative growth of the rate of introducing new gene names into the literature.

Thus, the data unequivocally supports the conclusion that, when the newly published scientific literature increasingly covered new proteins and their functions after 1945, this trend stopped at around 2000. A major growth of new literature thereafter is only detectable for proteins that are anyhow well studied (proteins with the T100 status moving into the T500 group) when, at the same time, research on non- or weakly studied protein-coding genes is increasingly abandoned.

3.3. Trends in the Discovery of New Non-Protein-Coding Gene and Non-Coding RNA Functions

From the viewpoint of the amount of available literature and new genomic items added per year, the research on non-protein-coding genes and non-coding RNAs is in its infancy (see **Figure 3**). A comparison with Figure 2 would place the non-coding RNA function discovery rates during the last two decades into similarity to the period before 1975 for the protein-coding genes. Until about 2000, the accumulation of literature per year for new non-coding RNAs previously not described was about constant at low values. Approximately ten new named entities were first mentioned every year (curve T0). For about a handful of new non-coding RNAs per year, sufficient FPEs were accumulated to cross the critical thresholds T1, T5, …, T20, and T25.

Only after 2000, the function discovery rates for non-coding RNAs started to develop some dynamism and strong growth. The mentioning of new non-coding RNA entities in the literature was at the level of 22 (T0), 9 (T1) and 7 (T5) in 2000. But it increased continuously to 386 (T0), 182 (T1) and 74 (T5) in 2017. For the medium thresholds (T10-T100), the values for 2017 were between 52 and 18 (with little changes during the last 3–5 years) when they were essentially zero (and always below 5) around 2000. Just singular cases of new non-coding RNAs make it above T500 every year even now.

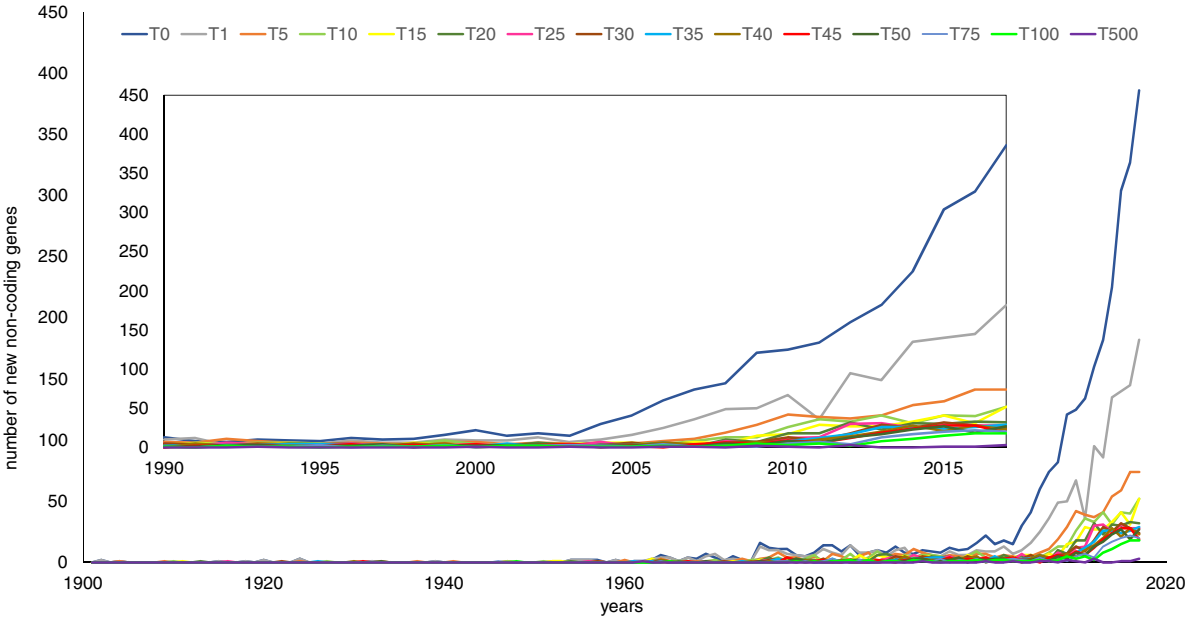

Figure 3. The number of new non-protein-coding genes in a given year with accumulated FPE score crossing different thresholds the first time. The notion "Tx" (where "x" is a natural number) is used to describe thresholds for a given named entity to cross the literature score threshold with at least "x" FPEs. So, "T0" denotes at least a single occurrence of the gene name (score>0) in the scientific literature. "T1" requires at least one full FPE to be accumulated. Accordingly, other thresholds such as T5, T10, T15, T20, T25, T30, T35, T40, T45, T50, T75, T100, and T500 are defined respectively. See Supporting Information file 2 for the exact gene numbers. The graph from 1990–2017 is zoomed on the left and shown in box.

One argument explaining the drop of attention in the literature towards new and scarcely studied protein entities might be the shifting towards research on functions of non-coding RNAs. To some extent, this is certainly the case as the magnitudes of change between 2000 and 2017 indicate. The values $\Delta Tx = Tx(2017)\text{-}Tx(2000)$ for protein coding genes are -340 (T0), -181 (T1), -83 (T5), -76 (T10), -29 (T15), 24 (T20), -40 (T25), 5 (T30), 43 (T35), -1 (T40), 11 (T45), 21 (T50), 28 (T75), 31 (T100), and 57 (T500). The respective data for non-protein-coding genes is 364 (T0), 173 (T1), 67 (T5), 48 (T10), 48 (T15), 31 (T20), 24 (T25), 23 (T30), 28 (T35), 23 (T40), 15 (T45), 25 (T50), 20 (T75), 17 (T100), and 2 (T500). Thus, the losses in protein function discovery for low thresholds (T0-T25) compare well with the gains in function discovery for non-coding RNAs in the same Tx range.

We illustrate the total function discovery dynamics from 1990 until 2017 in **Figure 4** (proteins and non-coding RNAs combined) and, for this data, we carried out a similar regression analysis as presented in Table 3. We find that the F-test supports a positive slope different from zero (i.e., significant growth with P-value<1%) only for T20 (3.0 additional function discoveries per year), T30 (3.6), T35 (3.8), T40 (3.0), T45 (2.8), T50 (2.6), T75 (3.5), T100 (3.4) and T500 (3.8). The P-value is > 0.22 for T0, T1, T5, T10 and T15 indicating a stagnate trend (with small slopes between -1.3 and 1.1 and the exception 2.3 for T0). Thus, annual combined protein and non-coding RNA function discovery rates taken together produce about a flat line over the years. Some growth can be seen in the categories T20-T50 and especially for T75, T100 and T500; thus, the well-studied targets receive even more attention. There is no sign of an overall function discovery rate boost after 2000. This is in strong contrast to the dynamics before the beginning of the new millennium.

4. Discussion

4.1. About the Levels of Darkness and Illumination in the Gene and Protein Function Space

The modern, scientifically based worldview requires understanding of structure and of mechanisms inherent to the matter so that rationally designed interventions produce predicted results. For biological systems, understanding of biomolecular mechanisms is critical before applications in medicine, biotechnology, agriculture and ecology will deliver.[1] For example, not surprisingly, cancers with known signaling pathways driving their growth can increasingly be well treated similar to infections whereas progress for those whose mechanism remain in the dark have not seen and will not get improvement in treatment success for years and decades to come.[29]

The first step towards understanding of biomolecular mechanisms is a full list of functions associated with every genomic region and gene.[2,3] Given the scale of world-wide life science research, it is surprising that, even for extensively studied simple model organisms at low evolutionary levels such as *Escherichia coli*[30,31] and *Saccharomyces cerevisiae*,[32] there are still lots of genes with unknown function.

To assess the completeness of discovery of functions encoded in the human genome, there should be first an agreement about how many functions are there in the code. The number of protein-coding genes in the human genome is an important aspect of this question; yet, despite of availability of its full DNA sequence (which should provide an upper threshold for the total number of encoded functions), the matter continues to be passionately discussed in the literature.[33,34] The number of protein–

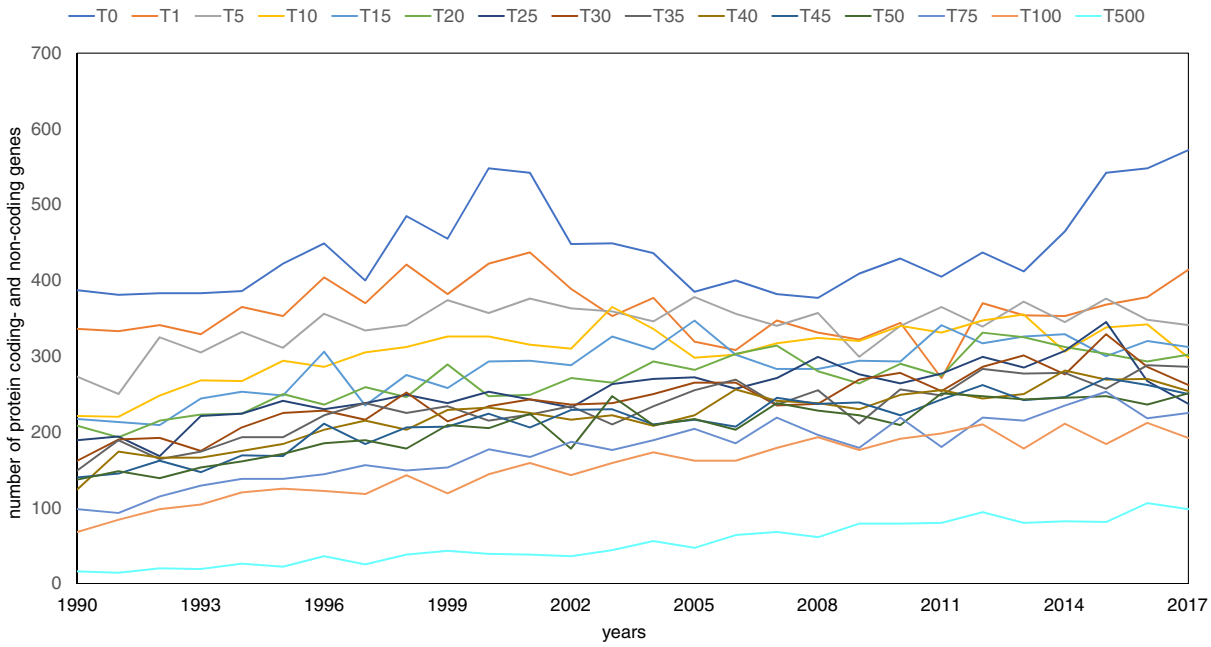

Figure 4. The number of any new genes in a given year with accumulated FPE score crossing different thresholds the first time for the period 1990–2017. Here, we show the total gene function discovery rate (combining the data for protein-coding genes and non-coding RNA) for the years 1990–2020. The notion "Tx" (where "x" is a natural number) is used to describe thresholds for a given named entity to cross the literature score threshold with at least "x" FPEs. So, "T0" denotes at least a single occurrence of the gene name (score > 0) in the scientific literature. "T1" requires at least one full FPE to be accumulated. Accordingly, other thresholds such as T5, T10, T15, T20, T25, T30, T35, T40, T45, T50, T75, T100, and T500 are defined respectively. See Supporting Information file 2 for the exact gene numbers.

coding genes changes with every genome release[35] and, especially from the side of small proteins, additions to the canonical human proteome have to be expected.[36–38] Besides small proteins, determining genes of rarely expressed proteins represents a formidable scientific challenge.

The estimated number of protein-coding genes in the literature is stated to be between 20 000 and 23 000.[25,36–38] Out of these, only 17 824 have ever been the target of a functional study as reported in the scientific literature until the end of 2017.

And a much smaller subgroup of elite 4817 proteins are quite well studied. Each of these targets has attracted ≥100 FPEs and, taken together, they represent 94.2% of all life science literature about genes and their functions (more than 4.2 million papers, patents, etc.). Thus, these targets exemplify the well illuminated part of the gene function space, typically with many aspects of their biological mechanism of action described. They are essentially the group of genes with target development levels T_{clin} (with known approved drugs) and T_{chem} (with known small molecular binders) as classified in.[8] It should be noted that, apparently, the event of full human genome sequencing in 2001 had negligible influence on function discovery among this group of elite genes. All of them had publications associated with them before the year of publishing the full human genome sequence (T0). Further, 4755 out of these 4817 genes (98.7%) have crossed their T10 threshold before 2001. About 2000 elite genes already had T500 status at 2001. So, these genes have typically been well known for a decade or longer at this time.

How many protein-coding genes remain in the dark? First, there is a group of 2200–5200 genes presumably encoded in the human genome that have never been studied functionally and not been mentioned in any paper (depending on the total estimate of proteins in the genome). We have to add those proteins whose functions are scarcely known. This is 1997 targets which did not even manage to cross T1 in 2017 and another 4571 below T10 (but above T1) in 2017. Together, this is easily a group of 10 000 protein-coding genes lost in the darkness of the function space with almost no illumination by any life science literature. There is yet another group of 6439 proteins in the limelight with 10–100 FPEs. To note, the matter of isoforms is not even included in this assessment.

To our knowledge, there has not been any estimate of the size of the function space for regions in the human genome for non-coding RNAs. So far, ≈120 non-coding RNAs attracted ≥100 FPEs and their functions can be considered well understood. Another ≈2500 named entities have been mentioned in the life science literature at various levels of scrutiny. Probably, this is just the tip of an iceberg with the by far larger majority of non-coding RNAs never been touched.

As a trend, annual new gene function discovery rates were growing in all categories of thresholds Tx in the years before 2000 (see Figures 2 and 3). Apparently, both the type of funding and the organizational structure of life science research as well as the mechanisms for selecting suitable individuals as principal investigators (PIs) were very appropriate for the medium-term, dedicated nature of work in this field of science. The spike at around 1975 is of special interest and we can only speculate about its origin here. Most likely, the reason is in the drastic expansion that the life science faculties and schools at universities have seen in

the late 60-ies and early 70-ies. Typically, the young professors have selected one new gene to study during their career and their first paper appeared at about 1975. Further, first nucleotide sequencing techniques have just been invented.[39]

The sectioning of time intervals chosen in table 3 could be made differently. Yet, we feel that, for the conclusions we are after in this work, the absolutely exact interval boundaries are not that critical. Our time sections roughly correspond to the adolescent, romantic era of life science research (until early 70-ies), the mature period (up to 2000) and the decay thereafter. The slopes are visibly different in the three time periods in Figure 2 whereas their exact numbers are not that important (but their orders of magnitude are).

Given the strong, robust and increasing growth of the protein function discovery rate before 2000, the more astounding is the decrease in protein function discovery after 2000 especially given the explosion of the total volume of life science literature. The number of annually added new entries in PUBMED[40] has grown from ~210 000 in 1970 to ~442 000 in 2000 and to ~870 000 in 2016.

At the other extreme of the publication universe, there are gene-superstars that attracted tens of thousands of FPEs. We did not explore the question why certain gene targets became research big names. Maybe, there are biological reasons for some cases (hubs in protein interaction networks, critical signaling entities in common pathologies, etc.) but, in many others, we think that quite profane factors such as being used as a biomarker for a process, the influence of some fashion in committees that allocate grants, etc. are the actual reasons. Of course, interacting genes have their FPEs grown in parallel by agglomeration.

4.2. When will we Understand the Human Genome?

This question was answered in 2012[2] based on a crude estimate of function discovery reports in top journals. The conclusion was that it will take us a hundred years just to complete the protein function catalogue. The more quantitative data in this work allows a better justified extrapolation and it leads to a slightly more optimistic result. The total amount of proteins receiving a first mentioning in 2017 was 186 and the number of proteins crossing T10 was 246. If these numbers do not drop further, the catalogue of estimated 10 000 missing protein functions might be completed in about 50 years.

The decrease in protein function discovery rates after 2000 had very substantial impact on the overall status of protein-coding gene annotation in the sequence databases. If discovery rates for protein functions had remained unchanged, the time for generating the same catalogue of functions would have taken, maybe, about half the years. With rates at the level of year 2000 just until 2017, >2800 more proteins had seen their first publication and additional >600 had crossed T10 in the same time.

The results of this literature analysis should be of concern at the side of science funders and administrators as it hints towards a waste of resources. When the event of full human genome sequencing was celebrated with pomp in 2001 and the availability of all gene sequences and their relative order in the genome was considered a major boost for the future development of science,[41] hardly anybody would have predicted that this event was the beginning of a sharp drop in the rate of protein function discovery (to just one third from 2000 towards 2017 if we take T0 as a measure).

Notably, there is no shortage of completely uncharacterized proteins to study. It also does not help to mention that this decrease should be imagined to be offset by function discoveries for non-coding RNAs. At best, the total rate of new function discovery has been stagnate when, at the same time, the international life science research community in academia and industry has rather expanded since 2000, total funding is at historically unprecedented heights and our research tools including the battery of omics methods is as good and as widely available as never before. Thus, it is unlikely to believe that cracking the problem of function for the remaining targets is so difficult that scientists at our time are incapable of being successful from the intellectual, resource or methodical points of view.

The most worrying point is that the fastest growing subgroup of protein targets attracting new life science literature is that of T500 proteins, those that are best studied anyhow when, at the same time, the rates for new T0, T1, T5 and T10 protein targets are rather declining. Thus, some part of the problems seem to be associated with the changes in science funding that occurred in the late 90-ies that direct people away from searching new functions encoded in the genome. It appears to be more aligned with the science system to propose research with incremental outcome in well-studied areas and to continue studying the same targets by inertia (as plentiful "preliminary data" is available). At the same time, efforts that aim at venturing into the really unknown are considered too risky or not sufficiently industrially applied by commissions assessing research grants and by scientists themselves. Reasons for scientists to systematically avoid big research questions have been discussed in detail elsewhere.[2] Most importantly, a function discovery is an effort of 5–10 years of a few scientific groups each costing about a million dollars per year. For most scientists and even PIs that have family and other long-term social obligations or that are in uncertain and short-term employment and funding conditions and need a fast success, it is not possible to unilaterally redirect their efforts towards such more fundamental issues.

The literature data does not support the claim that the availability of the full human genome sequence lead to a boost in biological function discovery associated with specific genomic regions or genes. Rather, the extremely successful trends for ever larger annual function discovery rates in 1980–1999 were reversed towards a sharp decline in reports of new protein functions. Even together with non-coding RNA function reports, the total discovery rate after 2000 is hovering at numbers that have already been seen around 2000. There is nothing comparable with the steep growth in the decade before. This is a paradoxical observation as the availability of the full genome sequence should and does ease the study of uncharacterized genomic regions.

The Battelle report from May 2011[42] finds that the human genome project had a great scientific impact implying considerable new functional insight besides a dramatic economic effect and the larger part of the report is dedicated to the former (pp. 17–52 of the main text in ref. 42). The argument is mainly qualitative (e.g., with Table 14 on page 23 of the main text about the biomedical applications of human genome sequencing) and with emphasis about the opportunities that the technology devel-

opment has opened. Especially, finding genomic aberrations in inherited diseases and determining their functional implications have become drastically easier compared with the previous era. Some credentials such as rational drug development are rather the results of independent successful developments in the preceding decade; full genome sequencing is not a prerequisite. It appears also that the event of full genome sequencing was perceived as a signal by the lay community to invest into academic life science research, biotechnology startups, and life science industry in general.

Maybe, we have to conclude that, because the support for classical function discovery work has dropped and been replaced by the preference toward hyped purely omics studies, many opportunities for function discovery associated with the full genome breakthrough have not been realized so far. Big funding organizations might revisit their own data and make conclusions. To note, omics methodologies just complement the biological research tool box available previously. They can nicely record expression patterns, subcellular localizations, etc. in a large scale. Yet, they do not substitute the old, proven techniques and approaches. Function of a gene is multifaceted and individual, it requires some dedicated, individual effort for every group of genes involved in a pathway or some type of biomolecular mechanism to discover the enzymatic changes, interactions, and signal transductions. Thus, an individual paper reporting about that gene's function is expected to appear at some time point but, alas, it is missing for thousands of genes.

Future fundamental research activities are desirably directed at function discovery of human genes that are not studied at all or to a very little extent. The group of conserved hypothetical proteins[43] in the human genome (some of them with homology up to the level of bacteria) is of special interest and it is most ridiculous that many of them have not been targeted by research efforts after so many decades. Yet, to compile a list of them that is accurate at this point of time is not trivial and, in this MS, we rather refrain from the matter. From the viewpoint of research efficiency, it might be commendable to concentrate efforts on simple organisms such as *Escherichia coli* or *Saccharomyces cerevisiae* to bring the list of non-annotated genes down to zero as experimentation with them is cheaper and easier than with the human system and any function discovery here informs about the function of other model organisms even if there is apparently no exact orthologue visible.

Further, the issue of alternative splicing variants, protein isoforms, etc. deserves much more attention; yet, at the level of today's literature mark-up, this is very difficult track in the literature. Yet, we can surely assume that, for the overwhelming number of gene targets, the issue of splicing variants and protein isoforms has never been touched in previous research.

5. Conclusions

The existing body of life science literature is misbalanced. A small group of <5000 elite proteins that have been well known since decades and <125 non-coding RNAs have attracted the overwhelming fraction of the life science literature dedicated to biomolecular mechanisms (94.2 and 76%, respectively). The trends in function discovery until 2000 worked toward improving the balance. At best, the biological function discovery rate for genomically defined named entities has not dropped (when it has gone down drastically for proteins) since 2001, the year of presenting the human genome sequence, but the expected boost in function discovery thereafter is suspiciously absent as there is no evidence for it in the life science literature.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors acknowledge financial support from A*STAR, the Novo Nordisk Foundation (NNF14CC0001), and the US National Institutes of Health (U54 CA189205 and U24 CA224370).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

complete human genome, gene function discovery, protein functions, scientific literature analysis

Received: July 30, 2018
Revised: September 7, 2018
Published online: October 30, 2018

[1] L. A. Levin, F. Behar-Cohen, *Trends Pharmacol. Sci.* **2017**, *38*, 1052.
[2] F. A. Eisenhaber, *J. Bioinform. Comput. Biol.* **2012**, *10*, 1271001.
[3] V. Kuznetsov, H. K. Lee, S. Maurer-Stroh, M. J. Molnar, S. Pongor, B. Eisenhaber, F. Eisenhaber, *Health Inf. Sci. Syst.* **2013**, *1*, 2.
[4] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, Fen. Eisenhaber, Y. M., *J. Mol. Biol.* **1998**, *283*, 707.
[5] B. Eisenhaber, S. Sinha, W. C. Wong, F. Eisenhaber, *Cell Cycle* **2018**, 1.
[6] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. Miklos, G. L., C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W.

Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, *Science* **2001**, *291*, 1304.

[7] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Headford, J. Howland, L. J. Kann, R. Lehoczky, P. LeVine, K. McEwan, J. McKeman, J. P. Meldrim, C. Mesirov, W. Miranda, J. Morris, C. Naylor, M. Raymond, R. Rosetti, A. Santos, C. Sheridan, N. Sougnez, N. Stange-Thomann, A. Stojanovic, D. Subramanian, J. Wyman, J. Rogers, R. Sulston, S. Ainscough, D. Beck, J. Bentley, C. Burton, N. Clee, A. Carter, R. Coulson, P. Daedman, A. Deloukas, I. Dunham, R. Dunham, L. Durbin, D. French, S. Grafham, T. Gregory, S. Hubbard, A. Humphray, M. Hunt, C. Jones, A. Lloyd, L. McMurray, S. Matthews, S. Mercer, J. C. Milne, A. Mullikin, R. Mungall, M. Plumb, R. Ross, S. Shownkeen, R. H. Sims, R. K. Waterston, L. W. Wilson, J. D. Hillier, M. A. McPherson, E. R. Marra, L. A. Mardis, A. T. Fulton, K. H. Chinwalla, W. R. Pepin, S. L. Gish, M. C. Chissoe, K. D. Wendl, T. L. Delehaunty, A. Miner, J. B. Delehaunty, L. L. Kramer, R. S. Cook, D. L. Fulton, P. J. Johnson, S. W. Minx, T. Clifton, E. Hawkins, P. Branscomb, P. Predki, S. Richardson, T. Wenning, N. Slezak, J. F. Doggett, A. Cheng, S. Olsen, C. Lucas, E. Elkin, M. Uberbacher, R. A. Frazier, D. M. Gibbs, S. E. Muzny, J. B. Scherer, E. J. Bouck, K. C. Sodergren, C. M. Worley, J. H. Rives, M. L. Gorrell, S. L. Metzker, R. S. Naylor, D. L. Kucherlapati, G. M. Nelson, Y. Weinstock, A. Sakaki, M. Fujiyama, T. Hattori, A. Yada, T. Toyoda, C. Itoh, H. Kawagoe, Y. Watanabe, T. Totoki, J. Taylor, R. Weissenbach, W. Heilig, F. Saurin, P. Artiguenave, T. Brottier, E. Bruls, C. Pelletier, P. Robert, D. R. Wincker, L. Smith, M. Doucette-Stamm, K. Rubenfield, H. M. Weinstock, J. Lee, A. Dubois, M. Rosenthal, G. Platzer, S. Nyakatura, A. Taudien, H. Rump, J. Yang, J. Yu, G. Wang, J. Huang, L. Gu, L. Hood, A. Rowen, S. Madan, R. W. Qin, N. A. Dvis, A. P. Federspiel, M. J. Abola, R. M. Proctor, J. Myers, M. Schmutz, J. Dickson, D. R. Grimwood, M. V. Cox, R. Olson, C. Kaul, N. Raymond, K. Shimizu, S. Kawasaki, G. A. Minoshima, M. Evans, R. Athanasiou, B. A. Schultz, F. Roe, H. Chen, J. Pan, H. Ramser, R. Lehrech, W. R. Reinhardt, M. McCombie, N. de la Bastide, H. Dedhia, K. Blöcker, G. Hornischer, R. Nordsiek, L. Agarwala, J. A. Aravind, A. Bailey, S. Bateman, E. Batzoglou, P. Birney, D. G. Bork, C. B. Brown, L. Burge, H. C. Cerutti, D. Chen, M. Church, R. R. Clamp, T. Copley, S. R. Doerks, E. E. Eddy, T. S. Eichler, J. Furey, J. G. Galagan, C. Gilbert, Y. Harmon, D. Hayashizaki, H. Hassler, K. Hermjakob, W. Hokamp, L. S. Jang, T. A. Johnson, S. Jones, A. Kasif, S. Kaspryzk, W. J. Kennedy, P. Kent, E. V. Kitts, I. Koonin, D. Krof, D. Kulp, T. M. Lancet, A. Lowe, T. McLysaght, J. V. Mikkelsen, N. Moran, V. J. Mulder, C. P. Pollara, G. Ponting, J. Schuler, G. Schultz, A. F. Slater, E. Smit, J. Stupka, D. Szustakowski, J. Thierry-Mieg, L. Thierry-Mieg, J. Wagner, R. Wallis, A. Wheeler, Y. I. Williams, K. H. Wolf, S. P. Wolfe, R. F. Yang, F. Yeh, M. S. Collins, J. Guyer, A. Peterson, K. A. Felsenfeld, A. Wetterstrand, M. J. Patrinos, P. Morgan, J. J. de Jong, K. Catanese, H. Osoegawa, S. Shizuya, Y. J. Choi, *Nature* **2001**, *409*, 860.

[8] T. I. Oprea, C. G. Bologa, S. Brunak, A. Campbell, G. N. Gan, A. Gaulton, S. M. Gomez, R. Guha, A. Hersey, J. Holmes, A. Jadhav, L. J. Jensen, G. L. Johnson, A. Karlson, A. Leach, R. A. Ma'ayan, A. Malovannaya, S. Mani, S. L. Mathias, M. T. McManus, T. F. Meehan, C. von Mering, D. Muthas, D. T. Nguyen, J. P.. Overington, G. Papadatos, J. Qin, C. Reich, B. L. Roth, S. C. Schürer, A. Simeonov, L. A. Sklar, N. Southall, S. Tomita, I. Tudose, O. Ursu, D. Vidovic, A. Waller, D. Westergaard, J. J. Yang, G. Zahoránszky-Köhalmi, *Nat. Rev. Drug Discov.* **2018**, *17*, 317.

[9] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, L. J. Jensen, *Nucleic Acids Res.* **2013**, *41*, D808.

[10] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, L. J. Jensen, *PLoS One* **2013**, *8*, e65390.

[11] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, C. von Mering, *Nucleic Acids Res.* **2015**, *43*, D447.

[12] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, C. von Mering, *Nucleic Acids Res.* **2017**, *45*, D362.

[13] A. Junge, J. C. Refsgaard, C. Garde, X. Pan, A. Santos, F. Alkan, C. Anthon, C. von Mering, C. T. Workman, L. J. Jensen, J. Gorodkin, *Database* **2017**, *2017*. https://doi.org/10.1093/database/baw167

[14] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, et al., *Nucleic Acids Res.* **2015**, *43*, D662.

[15] UniProt Consortium, *Nucleic Acids Res.* **2015**, *43*, D204.

[16] F. Eisenhaber, P. Bork, *Bioinformatics* **1999**, *15*, 528.

[17] D. Westergaard, H. H. Staerfeldt, C. Tonsberg, L. J. Jensen, S. Brunak, *PLoS Comput. Biol.* **2018**, *14*, e1005962.

[18] K. Kodukula, S. E. Maxwell, S. Udenfriend, *Methods Enzymol.* **1995**, *250*, 536.

[19] S. E. Maxwell, S. Ramalingam, L. D. Gerber, S. Udenfriend, *Proc. Natl. Acad. Sci. U. S. A* **1995**, *92*, 1550.

[20] S. E. Maxwell, S. Ramalingam, L. D. Gerber, L. Brink, S. Udenfriend, *J. Biol. Chem.* **1995**, *270*, 19576.

[21] M. Benghezal, A. Benachour, S. Rusconi, M. Aebi, A. Conzelmann, *EMBO J.* **1996**, *15*, 6575.

[22] D. Hamburger, M. Egerton, H. Riezman, *J. Cell Biol.* **1995**, *129*, 629.

[23] B. Eisenhaber, S. Eisenhaber, T. Y. Kwang, G. Gruber, F. Eisenhaber, *Cell Cycle* **2014**, *13*, 1912.

[24] E. Dolgin, *Nature* **2017**, *551*, 427.

[25] International Human Genome Sequencing Consortium, *Nature* **2004**, *431*, 931.

[26] N. Kresge, R. D. Simoni, R. L. Hill, *J. Biol. Chem.* **2005**, *280*, e3.

[27] H. A. Krebs, W. A. Johnson, *Biochem. J.* **1937**, *31*, 645.

[28] W. A. Engelhardt, M. N. Liubimova, *Mol. Biol.* **1994**, *28*, 1229.

[29] A. A. Leitner, A. Hochhaus, M. C. Müller, *Curr. Cancer Drug Targets.* **2011**, *11*, 31.
[30] K. Mukherjee, T. Narindoshvili, F. M. Raushel, *Biochemistry* **2018**, *57*, 2857.
[31] T. W. Ng, M. Ip, C. Y. H. Chao, J. W. Tang, K. P. Lai, S. C. Fu, W. T. Leung, *Appl. Microbiol. Biotechnol.* **2018**, *102*, 6257.
[32] L. Pena-Castillo, T. R. Hughes, *Genetics* **2007**, *176*, 7.
[33] Y. Hao, L. Zhang, Y. Niu, T. Cai, J. Luo, S. He, B. Zhang, D. Zhang, Y. Qin, F. Yang, R. Chen, *Brief. Bioinform.* **2017**.
[34] S. Samandi, A. V. Roy, V. Delcourt, J. F. Lucier, J. Gagnon, M. C. Beaudoin, B. Vanderperre, M. A. Breton, J. Motard, J. F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D. J. Hunting, A. A. Cohen, C. R. Landry, M. S. Scott, X. Roucou, *eLife* **2017**, *6*, e27860.
[35] F. L. Sirota, A. Batagov, G. Schneider, B. Eisenhaber, F. Eisenhaber, S. Maurer-Stroh, *J. Bioinform. Comput. Biol.* **2012**, *10*, 1250020.
[36] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, et al., *Nature* **2014**, *509*, 575.
[37] F. Abascal, D. Juan, I. Jungreis, L. Martinez, M. Rigau, J. M. Rodriguez, J. Vazquez, M. L. Tress, *Nucleic Acids Res.* **2018**.
[38] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, M. L. Tress, *Hum. Mol. Genet.* **2014**, *23*, 5866.
[39] W. A. Salser, *Annu. Rev. Biochem.* **1974**, *43*, 923.
[40] PUBMED. Detailed Indexing Statistics: 1965–2017. Completed and indexed citations, 7-9-2018, https://www.nlm.nih.gov/bsd/index_stats_comp.html
[41] F. S. Collins, V. A. McKusick, *JAMA* **2001**, *285*, 540.
[42] S. Martin, M. Grueber Economic impact of the human genome project, Battelle Memorial Institute, **2011**, https://www.battelle.org/docs/default-source/misc/battelle-2011-misc-economic-impact-human-genome-project.pdf
[43] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, *Nucleic Acids Res.* **2015**, *43*, D261.

